DOCUMENT RESUME

ED 314 494 TM 014 415

AUTHOR Benton, Roberta L.
TITLE The Statistical Power of Planned Comparisons.
PUB DATE Jan 90
NOTE 29p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, January 25-27, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Analysis of Variance; *Comparative Analysis; Factor Analysis; Research Design; *Statistical Analysis; *Test Use
IDENTIFIERS A Priori Tests; *Planned Comparisons; Post Hoc Tests; *Power (Statistics); Type II Errors

ABSTRACT

Basic principles underlying statistical power are examined; and issues pertaining to effect size, sample size, error variance, and significance level are highlighted via the use of specific hypothetical examples. Analysis of variance (ANOVA) and related methods remain popular, although other procedures sometimes have more statistical power against Type II error (not rejecting the null hypothesis when it is false). Most researchers recognize that when an ANOVA factor has three or more levels it may become necessary to conduct additional analyses to determine exactly where detected group differences actually arise. Many researchers approach this problem by conducting post hoc or unplanned tests if a statistically significant omnibus effect is isolated. A frequently superior alternative is to conduct planned tests. In addition to increasing the statistical power against Type II error, planned comparisons also force the researcher to think before selecting tests. These advantages can result in better research design and analysis. Sample data sets are used to illustrate the benefits of a priori or planned contrasts. The statistical power of planned comparisons is also explored using hypothetical factorial designs. Six tables illustrate the discussion. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED314494

sera.rno 1/11/90

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Roberta L. Benton

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The Statistical Power of Planned Comparisons

Roberta L. Benton

Louisiana State University Medical Center

Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX, January 25, 1990.

TM014415

BEST COPY AVAILABLE

ABSTRACT

Analysis of variance (ANOVA) and related methods remain popular, although other procedures sometimes have more statistical power against Type II error. Most researchers recognize that when an ANOVA factor has three or more levels it may become necessary to conduct additional analyses to determine exactly where detected group differences actually arise. Many researchers approach this problem by conducting post hoc or unplanned tests if a statistically significant omnibus effect is isolated. A frequently superior alternative is to conduct planned tests. In addition to increasing the statistical power against Type II error, planned comparisons also force the researcher to think before selecting tests. These advantages can result in better research design and analysis. Example data sets are used to illustrate the benefits of a priori or planned contrasts. The statistical power of planned comparisons is also explored utilizing hypothetical factorial designs.

Although the use of analysis of variance methods in educational research seems to be declining (Goodwin & Goodwin, 1985), ANOVA and ANCOVA remain the most frequently used methods for analyzing data (Elmore & Woehlke, 1988). However, even when OVA methods are appropriate, the use of a priori or planned comparisons may increase the statistical power of the analysis, thereby resulting in more valid research findings.

Such improvements may be very important. As Keppel (1982, p. 78) notes, "Systematic reviews of the psychological literature have revealed an astonishingly low level of power in the published experiments surveyed." Similarly, evaluation of tests of significance in reported research reveals that some researchers should be more concerned with the power of the test utilized (Gaither & Glorfeld, 1985; Rosnow & Rosenthal, 1989). Two problems can result from misapplication of parametric tests:

> First, the true significance level may be substantially different from the reported significance level. In other words, the researcher may be far more likely to reject a true hypothesis or draw an erroneous conclusion from the data than the actual reported level of Type I error would indicate. Second, the power of the test may be considerably lower than the researcher expects. This results in the inability to reject a false null hypothesis and

> the possibility of overlooking a result that could have practical implications. (Gaither & Glorfeld, 1985, p. 792)

The purpose of the present paper is to examine basic principles underlying statistical power. Issues pertaining to effect size, sample size, error variance, and significance level are made more concrete by utilizing specific hypothetical examples. The statistical power against Type II error of a priori or planned comparisons results in increased ability to detect differences among treatment conditions. This feature of planned comparisons is discussed and implications for research are suggested.

Statistical Power

As Minium and Clark (1982, p. 266) explain, "The power of a statistical test is the probability, given that Ho (the null hypotheses) is false, of obtaining sample results that will lead to its rejection... In other words, a powerful test is one that has a high probability of claiming that a difference exists when it really does." Cohen (1977), Keppel (1982), and Kirk (1968, 1984) offer similar explanations.

Power or sensitivity is dependent upon several factors including (a) the size of true treatment effects, (b) the sample size, (c) the degree of error variance, and (d) the significance level (Keppel, 1982; Kirk, 1968), and consideration of power should be among the first steps in planning any experiment (Keppel, 1982). The power of a test

is equal to one minus beta (where beta is the probability of a Type II error), and is determined by the four factors listed previously (Kirk, 1968). Anything the researcher does to increase power will by definition decrease the probability of a Type II error (Minium & Clark, 1982). "The smaller the Type II error (beta) the greater the power and, therefore, the greater the sensitivity of the test" (Keppel, 1982, p. 58) in detecting statistically significant differences.

As Keppel (1982, p. 59) notes, "Type II error may be decreased by adding to the number of observations in each treatment condition and by reducing the error variance through the design of a more precisely controlled experiment." Or, as Kirk (1968, p. 10) explains, "The two most common procedures for increasing power are 1) to increase the size of the sample, and 2) to employ an experimental design that provides a more precise estimate of treatment effects and a smaller error term."

The major source of error variance in behavioral science research is contributed by individual differences (Keppel, 1982), which can be partially controlled by the selection and assignment of subjects who are similar in their characteristics. Additionally, ANCOVA is purported to adjust error variance and estimates of treatment effects (Keppel, 1982), but such statistical controls have dangers that some researchers may not recognize (Thompson, 1988a). Prior to conducting research, the likely degree of error variance can

sometimes be estimated by consulting previous research (Keppel, 1982). The insights from such reviews may inform the researcher as to whether efforts to increase power may be especially necessary.

Expected effect sizes also bear upon such decisions, as noted previously. The effect size is determined by specific conditions of the investigation (Minium & Clark, 1982). The larger the difference between the means of the groups, the greater is the probability that a difference will be detected during significance testing. Thus, the larger the effect size the greater is the power of the test to detect a statistically significant result (Minium & Clark, 1982). Put simply, effect size involves the degree of relationship between the independent variable(s) and the dependent variable.

Table 1 presents the SPSS-X' output for a (3 X 2) factorial ANOVA with treatment effects of very large magnitude. The design employed a comparison between the lecture versus the computer assisted methods of instruction and three levels of IQ. Three subjects were assigned to each of the six groups, thus creating a balanced design (n=18). Results indicated zero effect sizes for omnibus main effects and nonsignificant results for both these independent variables, Method and IQ Group. All of the explained variance resulted from the interaction effect of IQGRP with METH (F=150, df=2/12, and $p< 0.05$), allowing rejection of the interaction null hypothesis. The eta squared effect size,

determined by calculating the ratio of the sum of squares (SOS) explained for interaction (300.0) to the SOS total (312.0), times 100 ((300/312)*100), was 96%, a very large effect size.

INSERT TABLE 1 ABOUT HERE.

These analyses reflect that, as Thompson (1988a, p. 5) notes, "Sample size is the primary influence on significance." As Minium and Clark (1982, p. 270) explain, "For any given effect size (other than zero) the larger the sample size the greater the power of the test", and that conversely, "smaller samples lead to a larger standard error and less power." It is important for the researcher to understand the relationship of both effect size and sample size to statistical significance testing. When a sample size is very large, even trivial effect sizes may be statistically significant (Thompson, 1989b).

The Table 2 computations illustrate these complex relationships between effect and sample sizes. The 3 X 2 ANOVA reported in Table 1 was manipulated to determine changes in significance test results associated with decreases in sample size. As reported in Table 2, with as few as n=7 subjects in this design, and a 96% effect size, the researcher would still be able to detect statistically significant results for the interaction test. In short, the larger the effect size, the smaller is the sample size required to reject

the null hypothesis. Rosnow and Rosenthal (1989) discuss the issue of power and significance testing and include a table giving the ratios of Type II to Type I error rates for various sample sizes, effect sizes and significance levels, which may be useful in designing studies with sufficient power for certain expected effect sizes.

INSERT TABLE 2 ABOUT HERE.

Utilizing Planned Comparisons to Increase Power

A hypothetical "one way" (Kerlinger, 1979, p. 84) design was devised to illustrate the superiority of planned comparisons over omnibus ANOVA testing, with respect to power considerations. The data set employed one independent variable with five levels. Ten subjects were assigned evenly, two to each of the five groups, thus creating a balanced design. The hypothetical dependent variable consisted of scores on a measure of performance.

Table 3 (Benton, 1989) presents the results of a classical omnibus ANOVA with the data. The procedure indicated no statistically significant differences among the means of the five groups (F=1.8, df=4/5, p=0.266). The eta-squared effect size, calculated by determining the ratio of the main effect sum of squares (SOS) to the total sum of squares, and then multiplying by 100, was 59% ((360/610) x 100).

INSERT TABLE 3 ABOUT HERE.

Although post hoc testing is done only when statistically significant results are obtained (Pedhazur, 1982), for illustrative purposes the data were subjected to seven commonly utilized a posteriori methods (LSD, Duncan, SNK, Tukey B, Tukey, LSDMOD, Scheffe). None of the post hoc or unplanned tests detected that any two group means were significantly different at the 0.05 level. Post hoc analyses were performed here to illustrate that even these procedures did not reveal the existing differences between means for the example data.

There are several reasons why planned comparisons or "ANOVA through regression" is superior to classical ANOVA. In classical OVA methods omnibus main effects are tested as a composite rather than being partitioned into separate components. One important advantage of planned comparisons is the protection that the planned comparison strategy provides against Type II error, i.e., the failure to detect true differences. In comparing a priori or planned versus post hoc or unplanned comparisons Kerlinger and Pedhazur (1973, p. 131) state:

> The tests of significance for a priori, or planned, comparisons are more powerful than those for post hoc comparisons. In other words, it is possible for a specific comparison

to be not significant when tested by post hoc methods but significant when tested by a priori methods.

As Thompson (1989b, p. 7) reiterates, "Regression approaches to ANOVA using a priori contrasts have more statistical power against Type II error."

A priori contrasts also help isolate specific sources of variance in the dependent variable. In the example analysis, all of the explained effect size demonstrated in the classical ANOVA was isolated solely from one planned contrast. Table 4 (Benton, 1989) presents the results of hand calculation of the keyout utilizing computer generated results from an "ANOVA through regression" and one contrast scheme. The resulting F ratio for the noteworthy contrast was statistically significant ($F=7.2$, $df=1/5$, $p<.05$). The effect size for the "A4" contrast was 59% ((360/610) x 100), as was the effect size for the omnibus main effect hypothesis in the classical ANOVA. The total explained variance was realized solely by the comparison of the mean of the two level 5 subjects with the dependent variable mean of the 8 remaining subjects in levels 1 through 4.

INSERT TABLE 4 ABOUT HERE.

Planned comparisons also require the researcher to think about relationships when designing proposed studies. Such planned contrasts (Rosnow & Rosenthal, 1989) encourage

precision of thought and theory, and "usually result in increased power and greater clarity of substantive interpretation" (p. 1281).

Statistical Power and the Factorial Design

A factorial design permits the manipulation of more than one independent variable in an experiment and the arrangement of treatment conditions allows determination of both separate and combined influences of each predictor variable (Keppel, 1982). By definition, all possible main effects and interaction combinations are tested when performing a factorial analysis.

The tests of standard factorial effects (main and interaction) qualify as planned comparisons by the very fact that the tests were planned before data were collected. Keppel (1982, p. 240) argues that, "Planning of simple effects and interaction comparison should be as common in designing a factorial experiment as are single-df comparisons in the designing of a single factor experiment."

Two issues related to statistical power are important to the researcher considering use of a factorial design. The first issue deals with Type I error and the possible increased tendency for making an "experimentwise" (termed "familywise" by Keppel, 1982) error when utilizing planned comparisons and a factorial design. Most researchers are completely unaware that the use of factorial analyses of balanced designs maximally inflates experimentwise error since (a) the maximum

number of tests are conducted, by definition, and (b) the omnibus tests are perfectly uncorrelated in balanced designs.

"Testwise" error rate is the probability of making a Type I error on a given hypothesis test. "Experimentwise" error rate involves the probability of making one or more Type I errors in the study as a whole. In studies in which more than one null hypothesis is tested the experimentwise Type I error rate is usually larger than the testwise error rate, as both Thompson (1987) and Fish (1988) explain.

The level of significance, or alpha, is the testwise probability of making a Type I error on a given significance test (Minium & Clark, 1982). The more we attempt to guard against Type I error by, for example, setting alpha at 0.01 instead of 0.05, the lower the power of the test becomes and the higher the likelihood of making a Type II error (not rejecting the null hypotheses when it is false) also becomes. Thus, attempts to guard against Type I error increase the chances for making a Type II error, if effect and sample sizes are fixed (Keppel, 1982). Setting alpha somewhere within 0.01 and 0.05 tends to give the researcher reasonable assurance that Ho (the null hypothesis) will not be rejected when it is true (Type I error) and at the same time usually does not inordinately raise the likelihood of making a Type II error (Minium & Clark, 1982).

But the use of factorial analysis increases the possibility of making a Type I error due to the fact that the

researcher is increasing the number of tests being conducted in the experiment. "Type I errors cumulate with each statistical test performed in a predictable and unavoidable fashion" (Keppel, 1982, p. 145). Keppel (1982) and Thompson (1988a) address both "testwise" (error per comparison) and "experimentwise" (error anywhere in the experiment) error, explaining how each is calculated and emphasizing the critical nature of the "experimentwise" error in the context of factorial design. Alarm is expressed concerning the potential for experimentwise Type I error due to the enormity of the total number of single-df comparisons possible for complex multiway factorial designs.

Fish (1988), Keppel, (1982), Kirk (1968), and Thompson (1988a), explain the method of computing the "experimentwise" error rate, which gives the researcher an estimate of the actual Type I error rate in an experiment in which a given number of comparisons are conducted. The formula for computing the level of significance for "experimentwise" error rate for orthogonal or uncorrelated contrasts is:

"experimentwise" error = $1 - (1-\text{alpha})^{k}$,

where k is the number of comparisons.

For the ensuing hypthotical (3 X 3 X 3) balanced factorial ANOVA analyses--where three main effects, three two way interactions, and one three way interaction are each tested at the 0.05 level of significance (the "testwise" error rate)

--the "experimentwise" error rate is calculated as 0.30. That is, 0.30 equals

$$1 - (1 - 0.05)^7 =$$

$$1 - (0.95)^7 =$$

$$1 - 0.70 = 0.30.$$

These results are for factorial analysis of omnibus tests; if all possible 1 df planned contrast comparisons (3 X 3 X 3 - 1 = 27 - 1 = 26) had been made, the inflation of experimentwise error would have been correspondibly larger.

Huberty and Morris (1988, p. 573) describe the influence of level of significance and effect size in assessing contrast effects:

> Jointly considering the two indicators, p and eta-squared, one can arrive at a conclusion regarding the existence of "real" (i.e., generalizable) contrast effects. Real effects exist if p is "small" and if eta-squared is "substantial." The determination of a small p value and a substantial eta-squared value is researcher- and situation-dependent. [Of course, the number of contrasts being investigated should be considered.] As in all of statistical inference, subjective judgment cannot be avoided. Neither can reasonableness! There are no general rules or set criteria for being reasonable.

Keppel (1982) advises the researcher to pre-plan for the proposed comparisons, the logic of the experimental design, and also to seek "a realistic estimate of power for the experiment in general and for the planned comparisons in particular" (p. 150.) Power estimates force the researcher to specify the smallest difference between two means that one wishes to detect and then to estimate the sample size necessary to achieve that power. Methods for estimating power are gives by Cohen (1977), and are illustrated by Minium and Clark (1982), Keppel (1982), and by others. Seeking power estimates forces the researcher to think about interrelationships among all the factors which influence power.

For the purposes of this paper additional hypothetical data sets were devised to illustrate situations where the factorial design may be advantageous to the researcher and also to illustrate when a factorial design could actually be a detriment to the study's outcome. In other words, a statistically significant result may not be obtained in some cases exactly <u>because</u> a factorial analysis--the default in several statistical packages--is used.

Say a hypothetical (2 X 2 X 2) factorial ANOVA was conducted to analyze the effects of lecture versus computer assisted instruction (CAI) with older and younger male and female subjects. Presume that the researcher only expects and

is solely interested in main effects. Table 5 presents the results of two three-way analyses.

If only the main effects are analyzed, no main effects are statistically significant. However, if a factorial analysis is conducted, the error sum of squares (SOS) is reduced by 20 (50-30), which will help make the mean square (MS) error smaller, which in turn makes all calculated F values for a fixed effects model larger. The degrees of freedom error get smaller by 4 (28-24), which will hurt by making the MS error larger, which in turn tends to make all calculated F values smaller. But for this particular case, the impact of the smaller error SOS outweighs the influence of the smaller df error, so on balance the factorial analysis helps to yield significance for the main effects.

INSERT TABLE 5 ABOUT HERE.

Conversely, it must be emphasized that factorial analyses are not always to the advantage of the researcher. Table 6 presents results from the same design hypothetically implemented with different subjects. In this case analyzing only main effects yields statistically significant results for all three main effects. However, no null hypotheses are rejected for the factorial analysis of the same data. In this case the factorial analysis hurts with respect to the main effect tests, because the help from a smaller SOS error (53-

50=3) is outweighed by the harm of a smaller df error (28-24=4).

INSERT TABLE 6 ABOUT HERE.

Summary

It must be noted that statistical "significance is not the end-all and be-all of research" (Thompson, 1988a, p. 18). Researchers often confuse statistical significance with the importance of the effect or with the likehood that the result is replicable, but significance testing actually evaluates neither of these two concerns. In addressing the dichotomous nature of significance testing decisions Rosnow and Rosenthal (1989), contend that "there is no sharp line between a 'significant' and a 'nonsignificant' difference; significance in statistics, like the significance of a value in the universe of values, varies continuously between extremes" (p. 1277).

The inherent value or importance of research is determined by what is being studied and the integrity and the quality of the design and implementation and interpretation of the study. Again, it is important that the researcher think carefully about the design of a proposed study and possible relationships among the variables in order to select the most appropriate statistical methods for the research. Four conclusions have been suggested in the previous discussion.

First, researchers should think carefully about design and the possible relationships among variables to select appropriate techniques that will increase statistical power against Type II error. This involves understanding of the complexities of effect size, statistical significance, error variance, and level of significance.

Second, when one or more ways in a design involve more than two levels, researchers should then consider using planned or a priori comparisons to help minimize chances of Type II error and to help locate specific sources of variance.

Third, researchers should evaluate the appropriateness of factorial designs for analysis of data. Such designs can dramatically inflate Type I error rates, but paradoxically can also result in the failure to detect statistical significance.

Fourth, researchers should understand the difference between statistical significance, magnitude of effect, and the importance of the research result. These distinctions have been described, but are also elaborated elsewhere (Carver, 1978; Cohen, 1977; Thompson, 1988c).

It should also be noted that the previous discussion involving univariate analyses also applies to the multivariate case, as well. As Swaminathan (1989, p. 231) argues,

> The often advocated procedure of following up the rejection of the null hypothesis with a more powerful multiple comparison procedure

> should be discouraged. First, the overall rejection of the null hypothesis does not guarantee that any meaningful contrast among the means will be significant... Second, since the overall test and the follow up procedure are unrelated, significant contrasts may be found even when the null hypothesis should not have been rejected. Third, follow up multiple comparison procedures which are unrelated to the overall test result in an inflation of the experiment-wise error rate. If multiple comparisons are of primary interest, a suitable multiple comparison procedure can be used without first performing an overall test.

In summary, in this paper some of the basic issues influencing statistical power have been examined. Discussion of important methodological considerations has been made more concrete by way of specific hypothetical data sets used as examples. For those who wish to pursue these various topics, the recent text by Keppel and Zedeck (1989) should prove helpful, as suggested by Thompson (in press).

References

Benton, R. (1989, November). Planned comparisons as better alternatives to ANOVA omnibus tests. Paper presented at the annual meeting of the Mid-South Educational Research Association, Little Rock. (ERIC Document Reproduction Service No. ED forthcoming)

Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review, 48(3), 378-399.

Cohen, J. S. (1977). Statistical power analysis for the behavioral sciences (rev. ed.). New York: Academic Press.

Elmore, P. B., & Woehlke, P.L. (1988). Statistical methods employed in American Educational Research Journal, Educational Researcher, and Review of Educational Research from 1978 to 1987. Educational Researcher, 17(9), 19-20.

Fish, L.J. (1988). Why multivariate methods are usually vital. Measurement and Evaluation in Counseling and Development, 21, 130-137.

Gaither, N., & Glorfeld, L. (1985). An evaluation of the use of tests of significance in organizational behavior research. Academy of Management Review, 10(4), 787-793.

Goodwin, L.D., & Goodwin, W.L. (1985). Statistical techniques in AERJ articles, 1979-1983: The preparation of graduate students to read the educational research literature. Educational Researcher, 14(2), 5-11.

Huberty, C. J., & Morris, J. D. (1988). A single contrast test procedure. Educational and Psychological Measurement, 48, 567-578.

Keppel, G. (1982). Design and analysis: A researcher's handbook (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Keppel, G., & Zedeck, S. (1989). Data analysis for research designs. New York: W.H. Freeman.

Kerlinger, F. N., & Pedhazur, E. J. (1973). Multiple regression in behavioral research. New York: Holt, Rinehart and Winston.

Kerlinger, F. N. (1979). Behavioral research: A conceptual approach. New York: Holt, Rinehart and Winston.

Kirk, R. E. (1968). Experimental design: Procedures for the behavioral sciences. Belmont, CA: Brooks/Cole.

Kirk, R. E. (1984). Elementary statistics. Monterey, CA: Brooks/Cole.

Minium, W. W., & Clarke, R. B. (1982). Elements of statistical reasoning. New York: John Wiley.

Pedhazur, E. J. (1982). Multiple regression in behavioral research (2nd ed.). New York: Holt, Rinehart and Winston.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44(10), 1276-1284.

Swaminathan, H. (1989). Interpreting the results of multivariate analysis of variance. In B. Thompson (Ed.),

Advances in Social Science Methodology (Vol. 1) (pp. 205-232). Greenwich, CT: JAI Press.

Thompson, B. (1987, November). The importance of a priori contrasts in analysis of variance research. Paper presented at the annual meeting of the Mid-South Educational Research Association, Mobile, AL. (ERIC Document Reproduction Service No. ED 292 835)

Thompson, B. (1988a, November). Common methodology mistakes in dissertations: Improving dissertation quality. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY. (ERIC Document Reproduction Service No. ED 301 595)

Thompson, B. (1988b). The importance of planned or focused comparisons in OVA research. Measurement and Evaluation in Counseling and Development, 21, 98-101.

Thompson, B. (1988c). A note about significance testing. Measurement and Evaluation in Counseling and Development, 20(4), 146-148.

Thompson, B. (1989a, January). Heuristics for understanding the concepts of interaction, polynomial trend, and the general linear model. Paper presented at the meeting of the Southwest Educational Research Association, Houston, TX. (ERIC Document Reproduction Service No. ED 304 453)

Thompson, B. (1989b). Statistical significance, result importance, and result generalizability: Three noteworthy

but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22(4), 2-5.

Thompson, B. (in press). Review of Data analysis for research designs by G. Keppel and S. Zedeck. Educational and Psychological Measurement.

Table 1

ANOVA (3 X 2)

Source	SOS	DF	MS	F	Sig F	Effect Size	Dec
Main							
Meth	0	1	0	0	1.0	0.00%	Not Rej
IQGRP	0	2	0	0	1.0	0.00%	Not Rej
2 Way Interaction							
Meth*IQGRP	300	2	150	150	0.0	96.15%	Rej
Residual	12	12	1				
Total	312	17	18.353				

Table 2
ANOVA (3 X 2)
Effect of Decrease in Sample Size to n=7

Source	SOS	DF	MS	F calc	F crit	Effect Size	Dec
Explained	**300**	**2**	**150**	150	3.88	**96%**	Rej
Residual	**12**	12	1				
Total	**312**	17					
Explained	**300**	**2**	**150**	75	5.14	**96%**	Rej
Residual	**12**	6	2				
Total	**312**	11					
Explained	**300**	**2**	**150**	25	19.00	**96%**	Rej
Residual	**12**	2	6				
Total	**312**	7					
Explained	**300**	**2**	**150**	12.5	200.00	**96%**	Not Rej
Residual	**12**	1	12				
Total	**312**	6					

Note. Entries in **bold** are presumed constant in estimating the effects of sample size changes. The main effects are presumed to remain zero, as in Table 1, and therefore are considered in df calculations (e.g., 17 - 1 - 2 - 2 = 12) but not in SOS or MS changes. Normally such analyses would be conducted only for possible balanced designs in which there were sufficient degrees of freedom total (at least 11) to test all factorial effects, i.e., n=12, 18, 24, and so forth. These analyses are reported only to give the reader an understanding of a procedure for evaluating sample size influences and to illustrate that a huge effect size (96%) generates significance even with a ridiculously small sample.

Table 3
Keyout for Classical ANOVA Analysis

Source	SOS	DF	Mean Square	F	Sig F	Effect Size
Explained	360	4	90.00	1.8	.266	59.00%
Residual	250	5	50.00			
Total	610	9	67.78			

Table 4
"ANOVA through Regression" Using Non-trend Contrasts

Source	SOS	DF	Mean Square	F	Sig F	Effect Size
A1	0	1	0.00	0.0		0.00%
A2	0	1	0.00	0.0		0.00%
A3	0	1	0.00	0.0		0.00%
A4	360	1	360.00	7.2	.005	59.00%
Residual	250	5	50.00			
Total	610	9	67.78			

Table 5
An Example of How Factorial Analysis can Help Yield Significance for Effects of Interest by Analyzing Even Effects Not of Interest

Nonfactorial analysis

Source	SOS	df	MS	Fcalc	Fcrit	Dec
Main						
Meth	**7.4**	**1**	**7.4**	4.144000	4.20	NS
Age	**7.0**	**1**	**7.0**	3.920000	4.20	NS
Sex	**6.0**	**1**	**6.0**	3.360000	4.20	NS
Residual	50.0	28	1.785714			
Total	**70.4**	**31**				

Factorial analysis

Source	SOS	df	MS	Fcalc	Fcrit	Dec
Main						
Meth	**7.4**	**1**	**7.4**	5.92	4.26	Rej
Age	**7.0**	**1**	**7.0**	5.6	4.26	Rej
Sex	**6.0**	**1**	**6.0**	4.8	4.26	Rej
2-Way						
Meth*Age	5.0	1	5.0	4.0	4.26	NS
Meth*Sex	5.0	1	5.0	4.0	4.26	NS
Age*Sex	5.0	1	5.0	4.0	4.26	NS
3-Way	5.0	1	5.0	4.0	4.26	NS
Residual	30.0	24	1.25			
Total	**70.4**	**31**				

Note. From Thompson (1990), with permission. Entries in **bold** remain constant.

Table 6
An Example of How Factorial Analysis can Hurt
by Yielding Nonsignificance for the Effects of Primary Interest

Nonfactorial analysis

Source	SOS	df	MS	Fcalc	Fcrit	Dec
Main						
Meth	**8.1**	**1**	**8.1**	4.279245	4.20	Rej
Age	**8.3**	**1**	**8.3**	4.384905	4.20	Rej
Sex	**8.0**	**1**	**8.0**	4.226415	4.20	Rej
Residual	53.0	28	1.892857			
Total	**77.4**	**31**				

Factorial analysis

Source	SOS	df	MS	Fcalc	Fcrit	Dec
Main						
Meth	**8.1**	**1**	**8.1**	3.888000	4.26	NS
Age	**8.3**	**1**	**8.3**	3.984000	4.26	NS
Sex	**8.0**	**1**	**8.0**	3.840000	4.26	NS
2-Way						
Meth*Age	0.5	1	0.5	0.240000	4.26	NS
Meth*Sex	1.0	1	1.0	0.480000	4.26	NS
Age*Sex	1.0	1	1.0	0.480000	4.26	NS
3-Way	0.5	1	0.5	0.240000	4.26	NS
Residual	50.0	24	2.083333			
Total	**77.4**	**31**				

Note. From Thompson (1990), with permission. Entries in **bold** remain constant.